MLPR
PROJECT

MULTI-LABEL MOVIE GENRE CLASSIFICATION

Table of CONTENTS

PROBLEM STATEMENT

Automated Classification of movie genres based on their plot synopsis

Enhancing Movie Recommendation Systems with Genre Classification

Motivation:
Users face information overload with vast movie choices. Genre classification aids in personalized movie selection. Accurate models open business opportunitie for companies in the entertainment industry.

Potential Applications:

1. **Recommendation Systems**: Improve suggestions across genres.
2. **Marketing**: Targeted campaigns for diverse audiences.

Impact:

1. **Enhanced Experience**: Personalized recommendations boost satisfaction.
2. **Efficient Management**: Streamline content organization.
3. **Market Insights**: Understand audience preferences for informed decisions.

Online Streaming Companies

Movie Review Websites

Flixster

Rotten Tomatoes

yahoo! movies

IMDb

LITERATURE Survey

Movie Label Classification of film Genres based on Synopsis

Jihadul Akbar , Ema Utami and Ainul Yaqin

https://www.researchgate.net/publication/369155718_Multi-Label_Classification_of_Film_Genres_Based_on_Synopsis_Using_Support_Vector_Machine_Logistic_Regression_and_Naive_Bayes_Algorithms

- Used Tf-Idf, bag of words, word2vec and doc2vec for feature extraction
- Label Powerset was used for problem transformation
- Used Support Vector Machine , Random Forest and XGBoost for model training
- Achieved f-1 score of 0.584 with SVM and Tf-idf

Our Take away : Feature extraction using TF-IDF and SVM as classifier and use them in combination (feature extraction should be done alongwith classifier)
Using class-weight parameter of SVM, could handle the imbalanced of classes(Assign higher weights to minority classes)

RESULTS

Algorithm	Matrix	Dataset		Dataset Mongolin et al [13]	
		TF-IDF	TF-IDF + Stemming	TF-IDF	TF-IDF + Stemming
SVM	*Accuracy*	0.140	0.142	0.122	0.127
	Precision	0.641	0.637	0.654	0.646
	Recall	0.536	0.546	0.510	0.527
	F1-Score	0.584	**0.589**	0.573	**0.580**
LR	*Accuracy*	0.085	0.095	0.081	0.091
	Precision	0.734	0.734	0.743	0.753
	Recall	0.286	0.313	0.226	0.258
	F1-Score	0.412	**0.439**	0.346	**0.384**
NB	*Accuracy*	0.104	0.107	0.098	0.102
	Precision	0.706	0.710	0.740	0.746
	Recall	0.373	0.370	0.308	0.310
	F1-Score	**0.488**	0.487	0.435	**0.438**

LITERATURE Survey

Analyzing Movies Using Phrase Mining

Daniel Lee, Huilai miao, Yuxuan Fan

https://www.researchgate.net/publication/369155718_Multi-Label_Classification_of_Film_Genres_Based_on_Synopsis_Using_Support_Vector_Machine_Logistic_Regression_and_Naive_Bayes_Algorithms

- Used Tf-Idf for word embedding
- Used Linear SVC as classifier and GridSearch for hyperparameter tuning
- Used Support Vector Machine , Random Forest and XGBoost for model training
- Evaluate the model using f-1 score
- F-1 score of 0.407 was achieved

Our Take away : GridSearch for hyperparameter tuning

Analyzing Movies Using Phrase Mining

Daniel Lee Huilai Miao Yuxuan Fan

March 7, 2021

Abstract

Movies are a rich source of human culture from which we can derive insight. Previous work addresses either a textual analysis of movie plots or the use of phrase mining for natural language processing, but not both. Here, we propose a novel analysis of movies by extracting key phrases from movie plot summaries using AutoPhrase, a phrase mining framework. Using these phrases, we analyze movies through 1) an exploratory data analysis that examines the progression of human culture over time, 2) the development and interpretation of a classification model that predicts movie genre, and 3) the development and interpretation of a clustering model that clusters movies. We see that this application of phrase mining to movie plots provides a unique and valuable insight into human culture while remaining accessible to a general audience, e.g., history and anthropology non-experts.

DATASET OVERVIEW

Collection:

- The dataset was sourced from Kaggle, originally compiled from IMDB.
- The primary motivation for selecting this dataset was its high dimensionality, aiming to enhance data accuracy and optimize model performance.

Dataset Details:

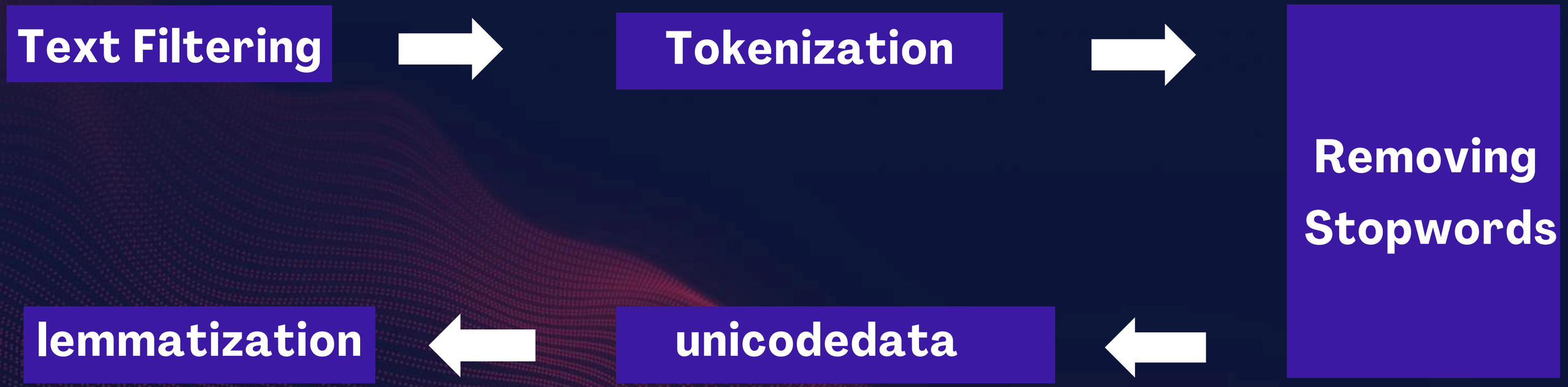
- 117194 rows and 4 columns constitute the dataset.
- There are no null values present in any of the entries.
- The target feature for our analysis is 'genre'

```
df.head(20)
```

	title	plot	plot_lang	Genre
0	"#7DaysLater" (2013)	dayslat interact comedi seri featur ensembl ca...	en	Comedy
1	"#BlackLove" (2015) {Crash the Party (#1.9)}	week leave workshop women consid idea ladi stu...	en	Reality-TV
2	"#BlackLove" (2015) {Making Lemonade Out of Le...	women start make stride toward find version ha...	en	Reality-TV
3	"#BlackLove" (2015) {Miss Independent (#1.5)}	women independ strong becaus theyv face strife...	en	Reality-TV
4	"#BlackLove" (2015) {Sealing the Deal (#1.10)}	despit go life chang process past week women s...	en	Reality-TV
5	"#Cake" (2015)	cake hour long serial narrat comedi manhunt hi...	en	Comedy
6	"#CandidlyNicole" (2013) {What's My Sports Vib...	marri sport nut sometim help know littl favori...	en	Reality-TV
7	"#Elmira" (2014)	elmira follow stori bunch stranger respond cra...	en	Comedy
8	"#Hashtag: The Series" (2013)	friend follow like fall hashtag follow love li...	en	Comedy
9	"#LawstinWoods" (2013)	lawstinwood follow stori stranger take live pl...	en	Sci-Fi
10	"#LawstinWoods" (2013) {The Case of the Twins ...	guy wake unfamiliar wood come girl claim wake ...	en	Sci-Fi
11	"#LawstinWoods" (2013) {The Happening (#1.3)}	gang discuss shock realiz ident twin didnt kno...	en	Sci-Fi
12	"#MonologueWars" (2014)	monologu war pit side ensembl friend contest h...	en	Drama
13	"#NotMadMonday" (2015)	notmadmonday new fast pace talk show star unli...	en	Comedy
14	"#SmurTv" (2016)	sketch comedi creator onlin seri swaggapuss ge...	en	Comedy
15	"#SpongeyLeaks" (2016)	trevor aunt barbara eyster live multipl sclero...	en	Biography, Crime, Reality-TV
16	"#TanCosmo" (2014)	monica fonseca colombian host help creat style...	en	Talk-Show
17	"#VanLifeAttila" (2016) {An Ideal Life (#3.5)}	attila talk emot futur depress media affect re...	en	Adventure, Biography, Comedy, Drama, History
18	"#mykpop" (2013)	global phenomenon k pop wave live fanat world ...	en	Music, Reality-TV
19	"#mykpop" (2013) {(#1.2)}	k pop grow check mnet america newest docu real...	en	Music, Reality-TV

Note: There were no ethical concerns as such when collecting data from both IMDB(by author) and from kaggle(by us).

Data Preprocessing



Text Filtering:

- Remove Punctuation Marks ,HTML tags, numerical values, and non-English texts.

Tokenization:

- Break down the text stream into individual tokens.
- Example: "that was part of a major crackdown" becomes ["that", "was", "part", "of", "a", "major", "crackdown"].

Remove Stopwords:

- Eliminate generic words with little semantic relevance from the text.
- Utilize pre-defined stopword lists.
- Common stopwords include "and," "are," "this," etc.

Unicodedata:

- Convert text to a standard form so that different representations of the same characters are treated as identical.
- For example character like "é" and "e" are treated as identical.

Lemmatization:

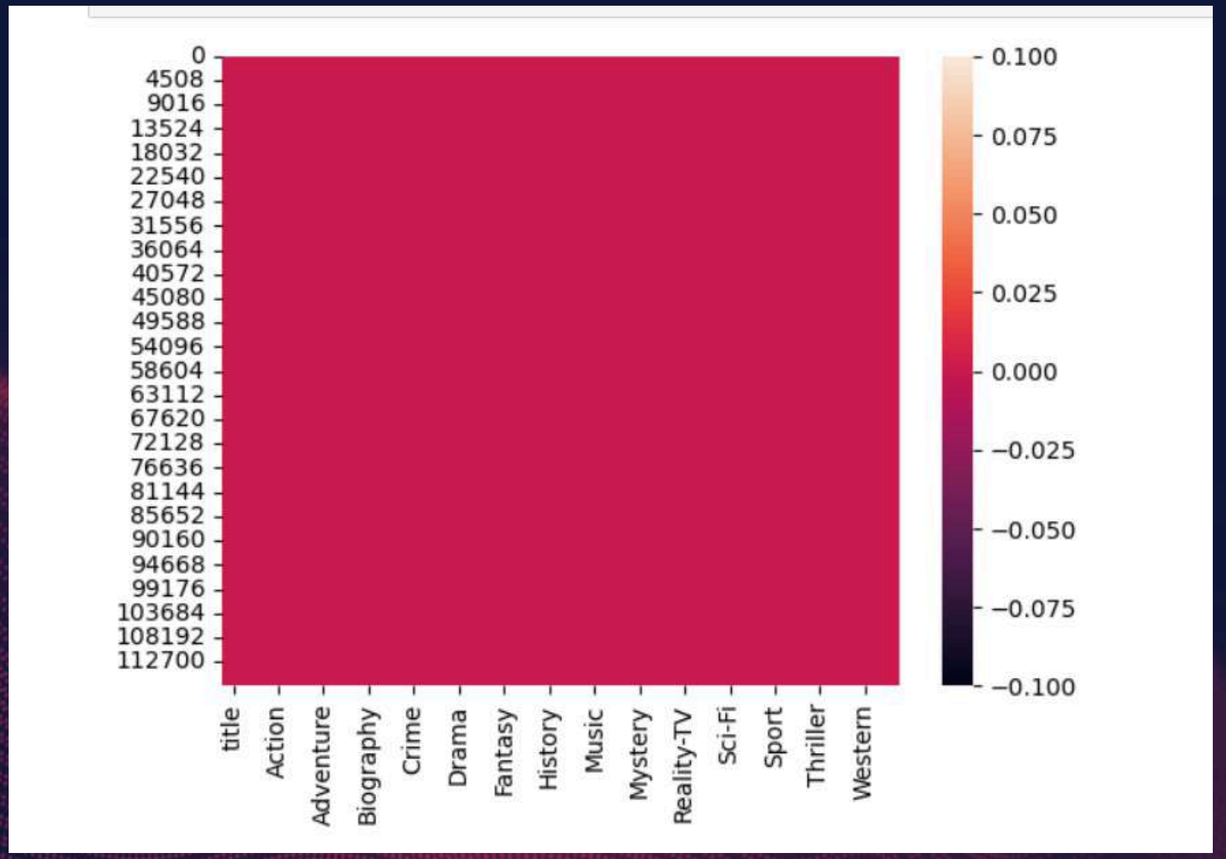
- Reduce words to their base or root form.
- Helps in achieving more meaningful and accurate text analysis by grouping together different forms of the same word (e.g., "running," and "runs" are all lemmatized to "run").

Exploratory Data Analysis

Missing Data Fields

The dataset shown below exhibits no missing values across its features.

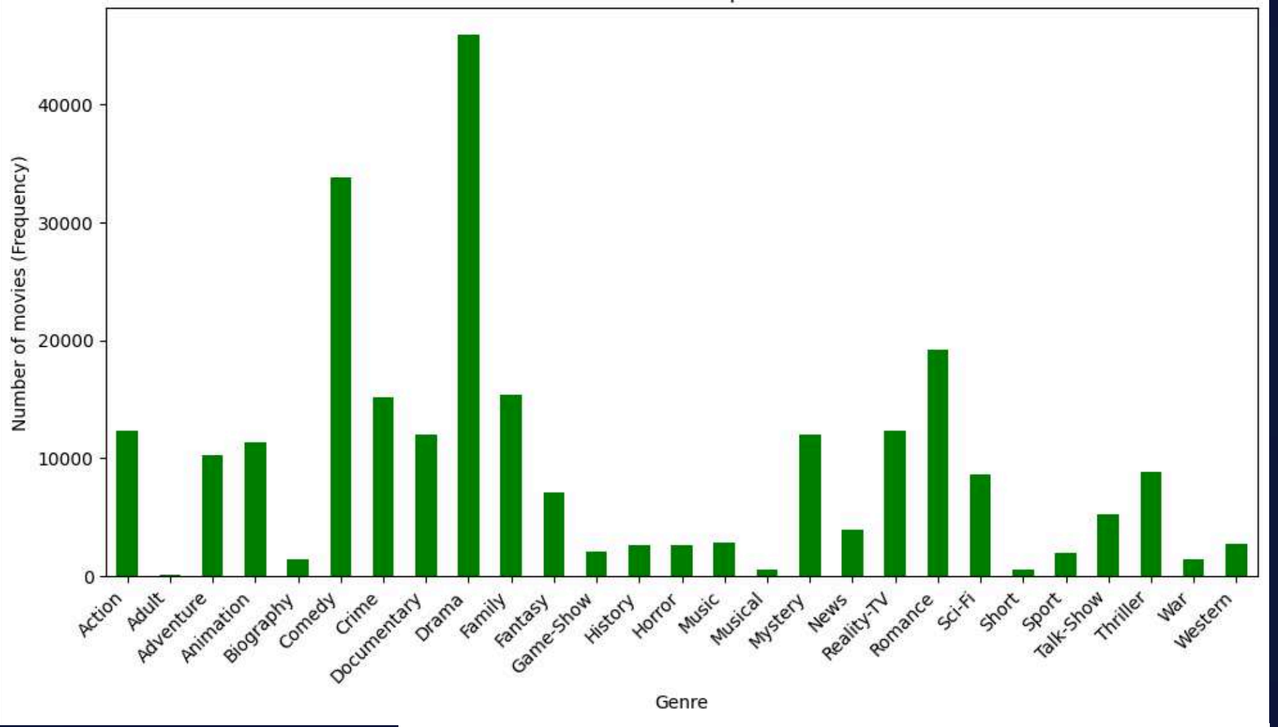
```
In [43]: (mydata.isnull().sum()/mydata.shape[0])*100

Out[43]: title          0.0
         plot           0.0
         Action         0.0
         Adult          0.0
         Adventure      0.0
         Animation      0.0
         Biography      0.0
         Comedy         0.0
         Crime          0.0
         Documentary    0.0
         Drama          0.0
         Family         0.0
         Fantasy        0.0
         Game-Show      0.0
         History        0.0
         Horror         0.0
         Music          0.0
         Musical        0.0
         Mystery        0.0
         News           0.0
         Reality-TV     0.0
         Romance        0.0
         Sci-Fi         0.0
         Short          0.0
         Sport          0.0
         Talk-Show      0.0
         Thriller       0.0
         War            0.0
         Western        0.0
         plot_lang      0.0
         dtype: float64
```



Number of Movies per Genre

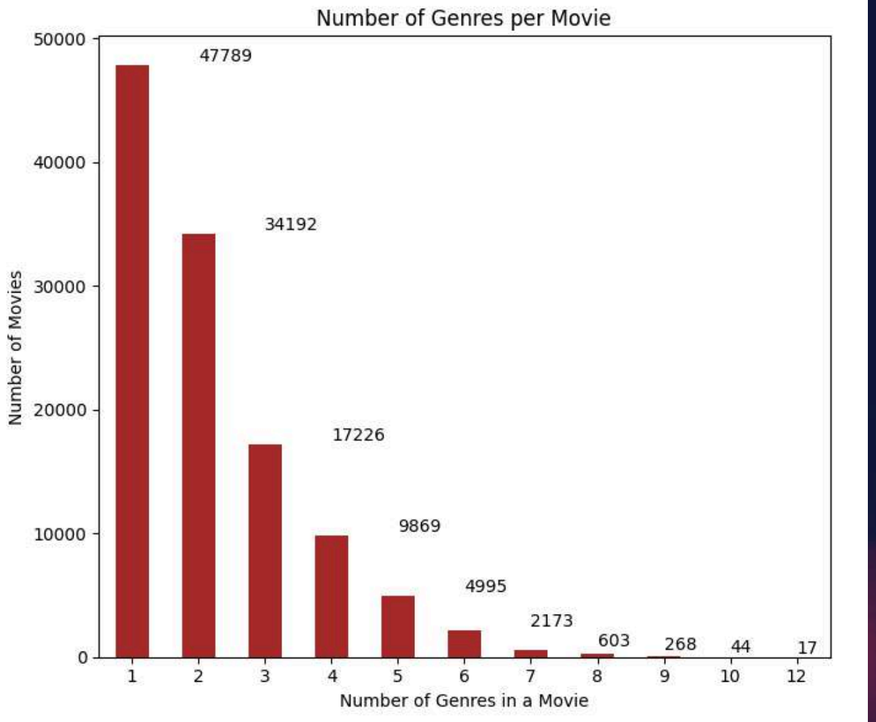
Observations:

- **Adult genre: 61 movies (lowest)**
- **Drama genre: 45,891 movies (highest)**
- **Comedy genre: 33,870 movies(second highest)**



Number of Genres per Movie

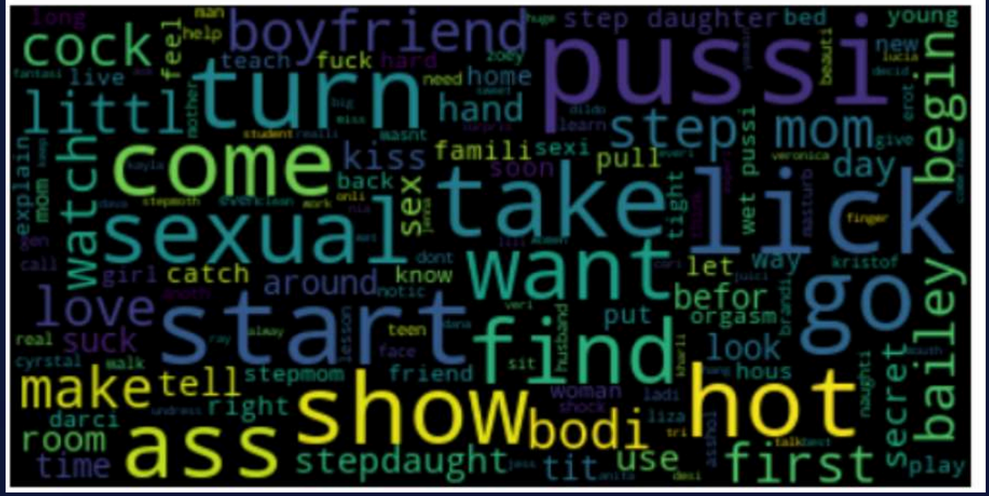
- **Number of genres each movie is classified into**
- **The majority of movies are categorized with 1 or 2 genres.**
- **On average, each movie is associated with 2.5 genres.**
- **18 movies are classified using 12 genres.**



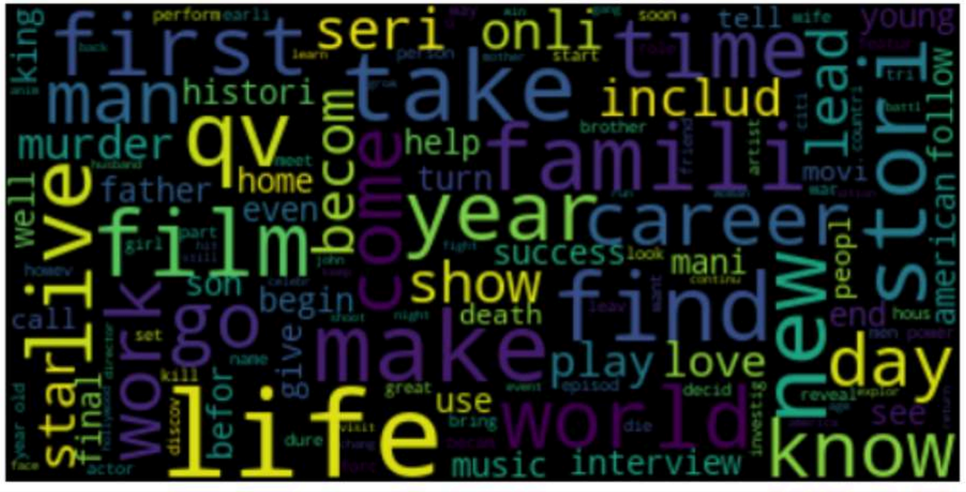
Word Cloud plots

Movie Genre: Action

Movie Genre: Adult



Movie Genre: Biography



Movie Genre: Comedy

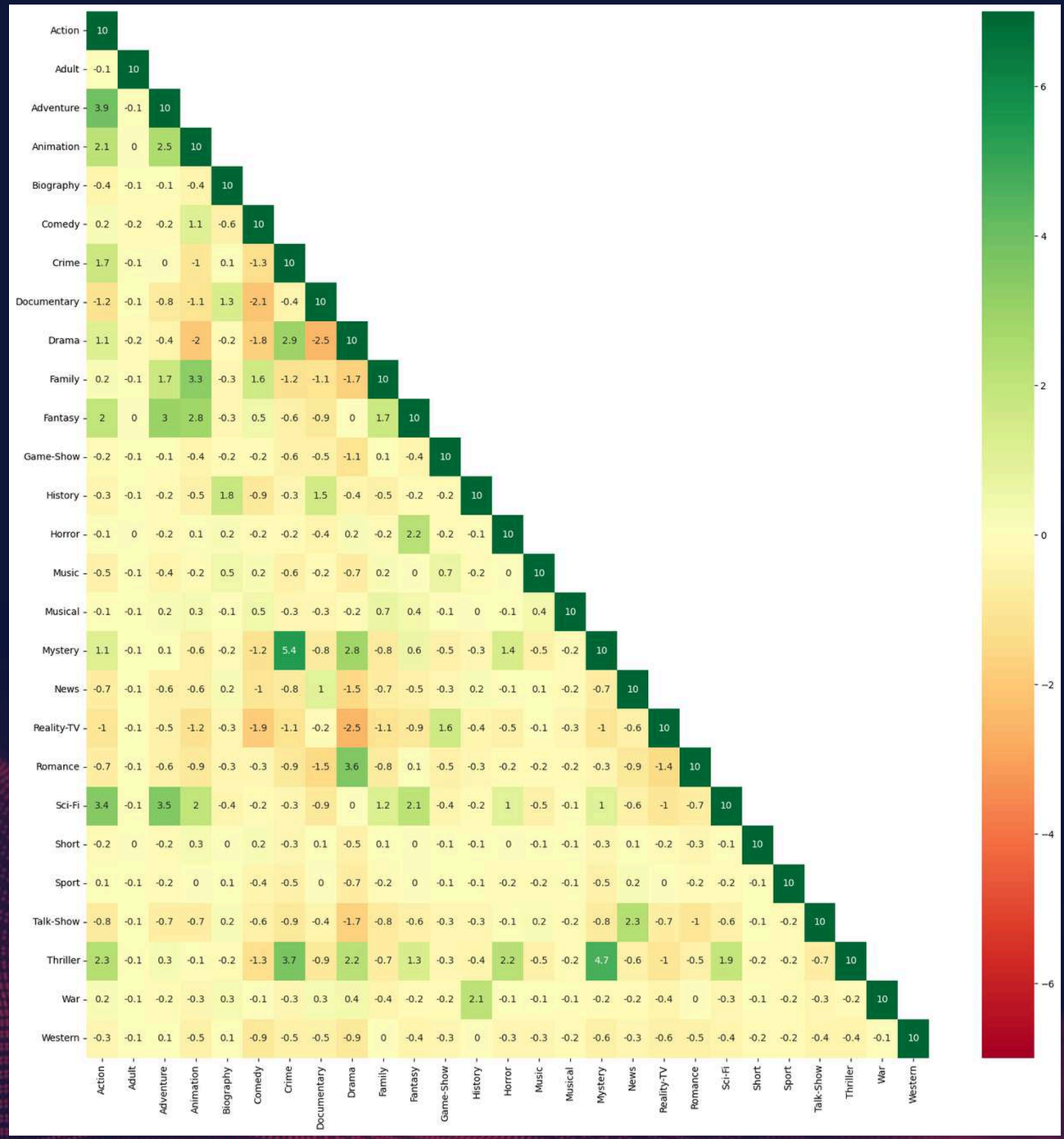
Correlation matrix

Genres with strong positive correlation:

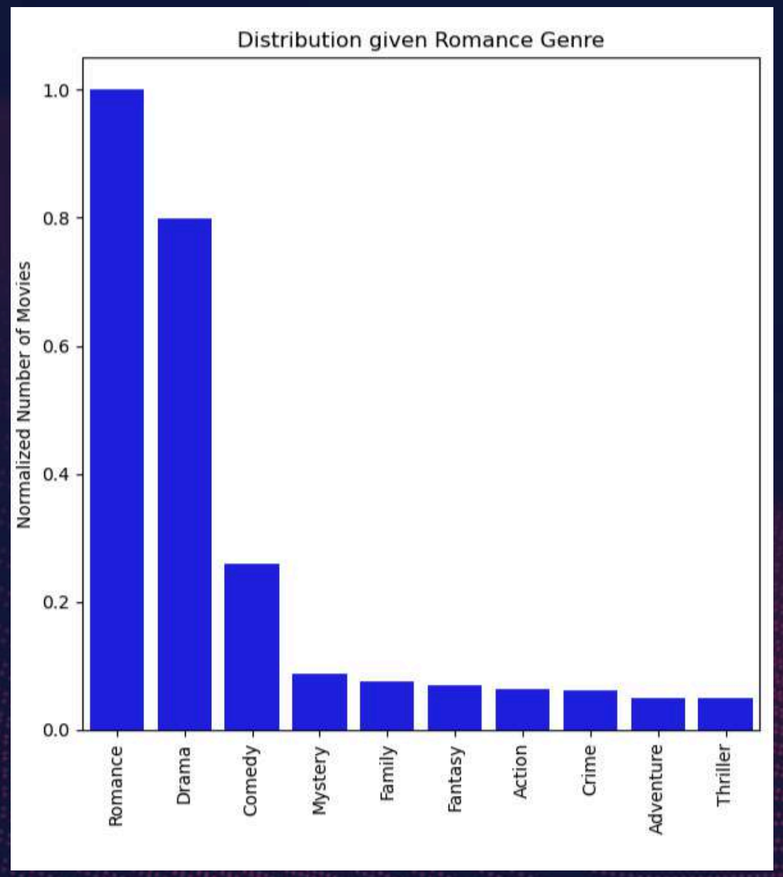
- Action, Adventure, & Sci-Fi
- Animation, Fantasy, & Family
- Crime, Thriller, Mystery, & Drama
- Biography, Documentary, & History
- Drama & Romance
- Game-show & Reality-TV
- Horror, Thriller, & Fantasy
- Talk-show & News
- War & History

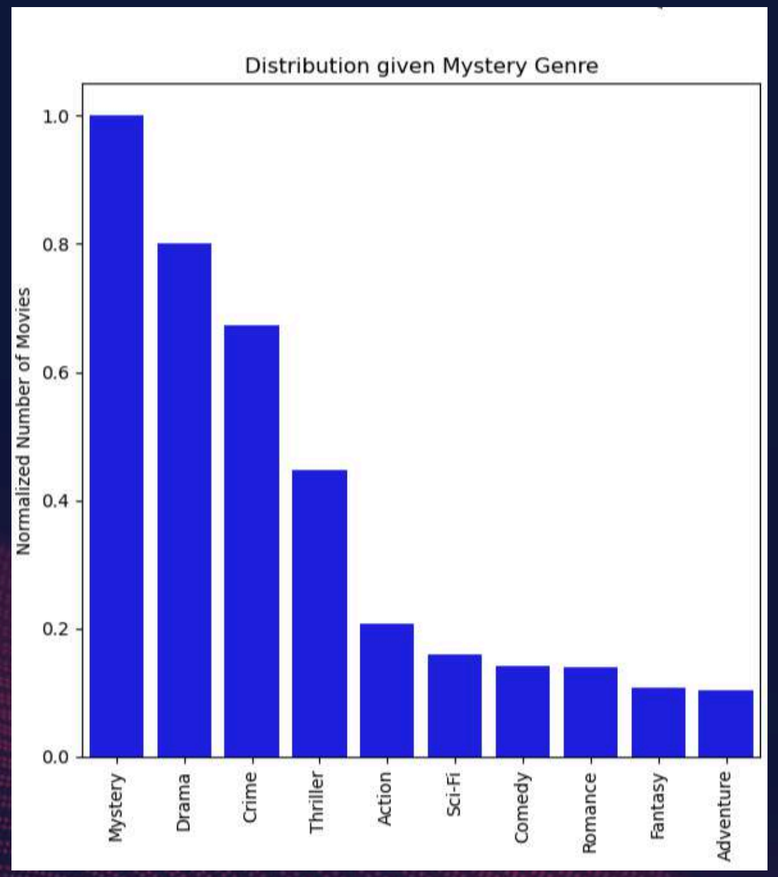
Genres with strong negative correlation:

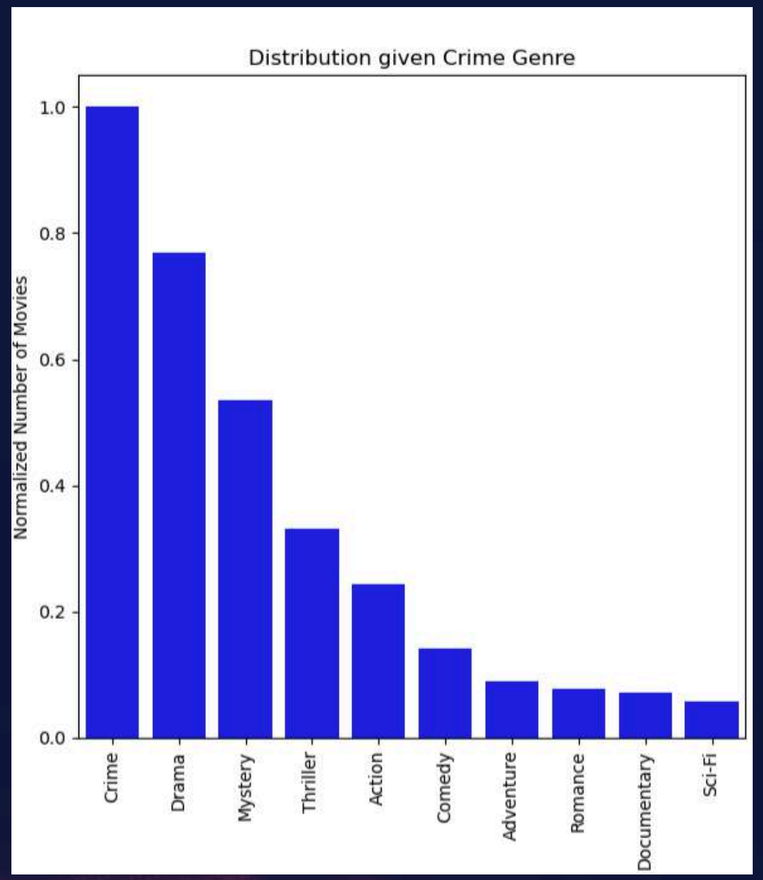
- Animation & Drama
- Comedy with Documentary & Reality-TV
- Documentary with Comedy, Drama, & Romance
- Drama with Animation, Reality-TV, & Comedy



Multi-Genre Distribution Plots





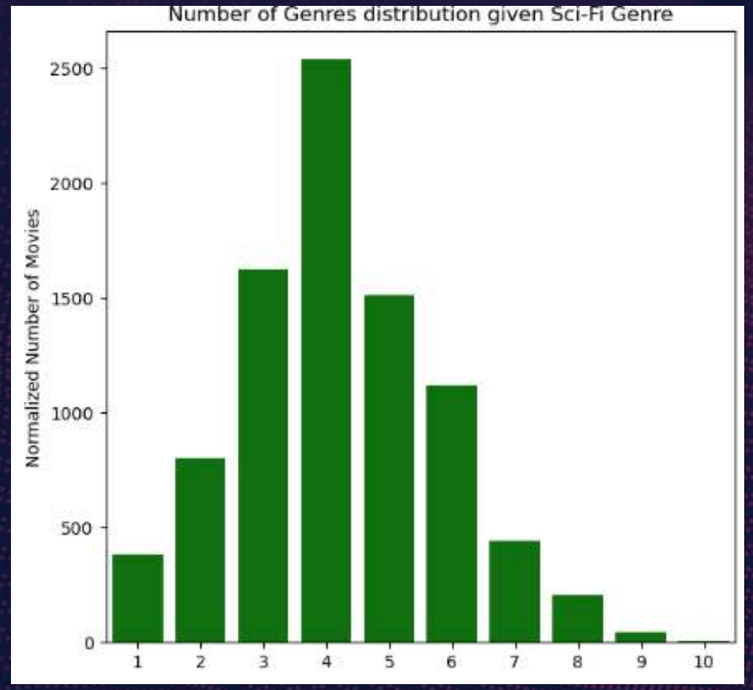


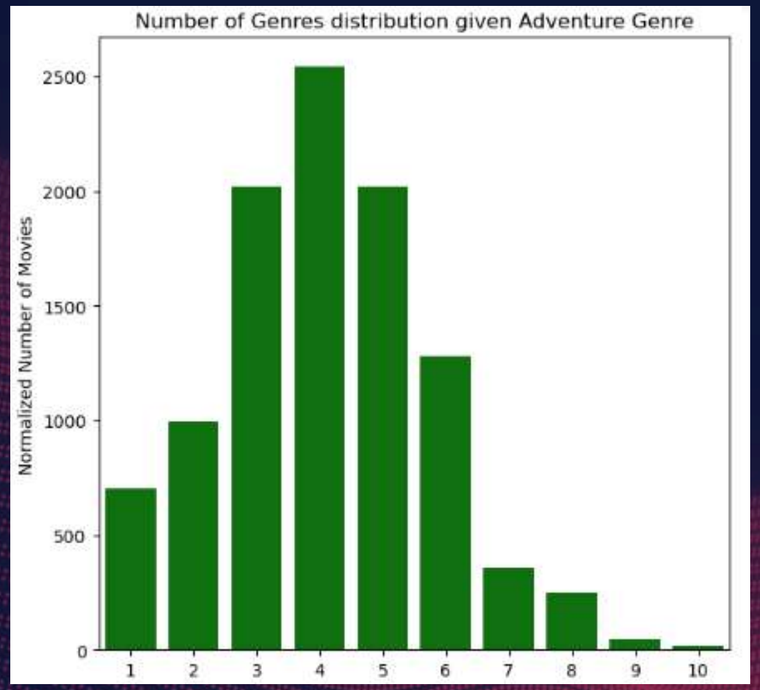
- **Here we try to see if a movie belongs to a certain Genre, what are the other Genres it might fall under.**
- **From the above plots, we can see 80% of Romance , Mystery and Crime movies also fall in Drama Genre**

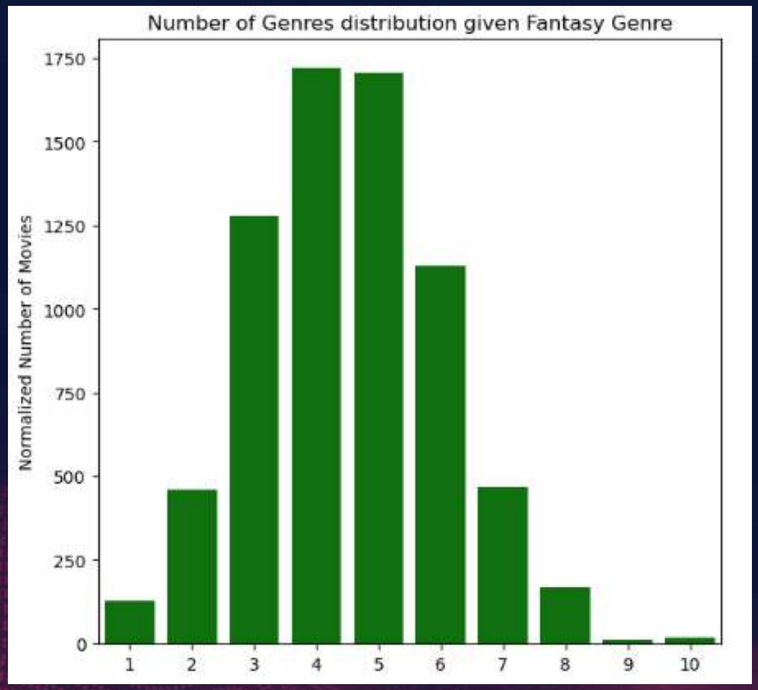
Number of Genres per Movie

- **Here we try to see how many genres each movie is classified into for each of the 27 genres.**

- **Most of the Sci-Fi , Adventure ,Fantasy, Action, Animation, Horror and Thriller movies have 3 to 6 categories**





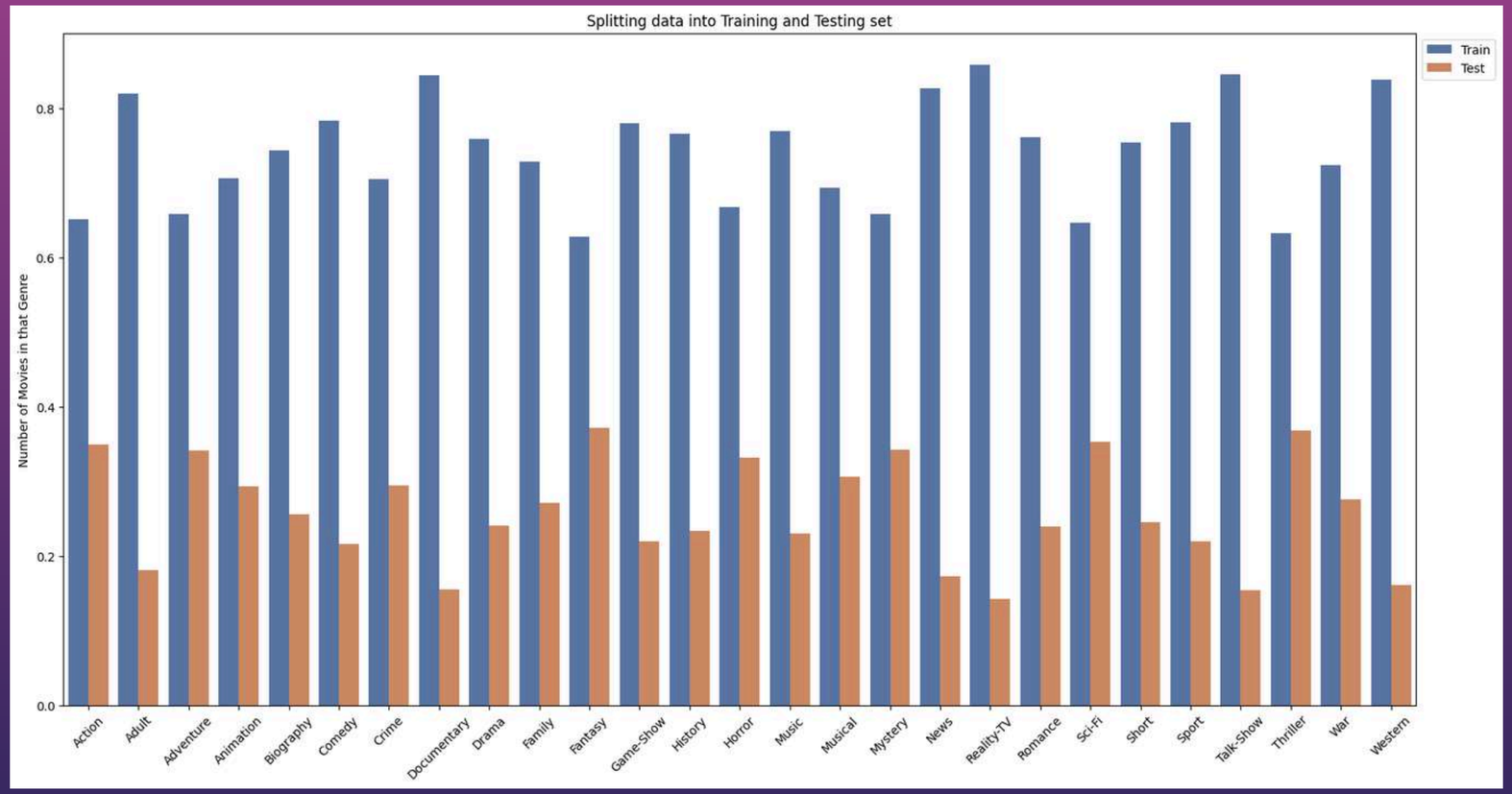


Modelling

Train/Test Split

Imbalanced Data with genre occurrences ranging from 61 (Adult) to 45891 (Drama)

- Split the provided data set ensuring that both the data sets have a minimum fraction (~0.2) of every label.
- looping through each category and include 0.2 fraction of that category into the test data set. Clearly at the end of the loop, the number of occurrences of each category will be greater than 0.2 fraction since most movies that are being included into the test set as a genre also are categorized with other genres.



Train/Test split such that

At least 80% of the samples in the training set At least 20% of the samples in the test set

Our Methodology to solve the multilabel classification problem

Text Encoder

Transform the words from the content to numerical vector
so that algorithm can work upon

- TF- IDF - Enhances the importance of words unique to a document by assigning higher weights to less frequent words across the corpus, making them more significant. It combines two metrics,'term frequency' and 'inverse document frequency'.
- CountVectorizer- Unlike TF-IDF which measures the importance of words based on their document frequency, CountVectorizer simply counts the number of times each word appears in a document.

Problem Transformation(Algorithm used)

- Transforming multi-label problem into single label problem
- This method is carried out using **Binary Relevance**
- Simple technique which basically treats each label as a single class)

Classifiers used

- LinearSVC
- Multinomial Naive Bayes
- Logistic Regression

MODELS

We are training four models with different combination of algorithms, encoders and classifiers in order to get the best model.
We are generating the f1 score for all the four models and observing which model has the highest score.

Text Encoder	**Problem Transformation (algorithm)**	**Classification Models**
TF-Idf	Binary Relevance	MNB
TF-Idf	Binary Relevance	Logistic Regression
TF-Idf	Binary Relevance	Linear SVC
Count Vectorizer	Binary Relevance	Linear SVC

Reasons for selecting these algorithms:

1. **Logistic Regression**
- Probabilistic Output: Logistic regression naturally provides probabilistic output, allowing for easy threshold adjustment and probabilistic interpretation of predictions.
- Efficiency: Logistic regression is computationally efficient, making it suitable for large-scale datasets
- Robustness to Irrelevant Features: Logistic regression is robust to irrelevant features, as it tends to assign low weights to irrelevant features during model training.

2. **Multinomial Naive bayes**
- Multinomial naive Bayes has few hyperparameters and is relatively insensitive to their values, reducing the need for extensive hyperparameter tuning and simplifying the model selection process.
- Multinomial naive Bayes naturally extends to multi-class (and multi-label) classification tasks, allowing for straightforward modeling of genre combinations.

3. **Linear SVC**
- Handling Class Imbalance: By assigning higher weights to minority classes and lower weights to majority classes, the SVC can effectively mitigate the impact of class imbalance and improve the model's performance on minority classes.
- Linear SVC aims to maximize the margin between classes, leading to better generalization performance and robustness to noise

Building the Base

Base1= TF_IDF + Binary Relevance

Base2= CV + Binary Relevance

. TF-IDF Vectorization:

- TfidfVectorizer is applied within a Pipeline to the plot column of the movie data.
- Transforms textual content plot column into a numeric form.
- Parameters Used: max_df, min_df, and ngram_range are tuned to optimize this transformation.
- For example, max_df=0.5 means ignoring terms that appear in more than 50% of the documents.

. Count Vectorizer(CV):

- CountVectorizer counts the number of times each word appears in a document.
- Paramaeters used are same as TF-IDF

Binary relevance

- The multi-label classification problem is decomposed into multiple binary classification tasks, one for each label.
- Used in multi-label classification where each label is treated as a separate binary classification problem.
- For each label, a separate classifier (e.g., logistic regression, Naive Bayes) is trained independently to predict whether that label should be assigned to an instance or not.

X	Y_1	Y_2	Y_3	Y_4
$X^{(1)}$	0	0	0	1
$X^{(2)}$	1	0	0	0
$X^{(3)}$	1	0	0	1
$X^{(4)}$	0	1	0	0
$X^{(5)}$	0	1	1	0

X	Y_1
$X^{(1)}$	0
$X^{(2)}$	1
$X^{(3)}$	1
$X^{(4)}$	0
$X^{(5)}$	0

X	Y_2
$X^{(1)}$	0
$X^{(2)}$	0
$X^{(3)}$	0
$X^{(4)}$	1
$X^{(5)}$	1

X	Y_3
$X^{(1)}$	0
$X^{(2)}$	0
$X^{(3)}$	0
$X^{(4)}$	0
$X^{(5)}$	1

(**Base to Classifires**) **PROCESS** **Trained model**

Training Process:

- Pipeline: We have created a pipeline which has two components ‘TfidfVectorizer()’ -> for feature extraction and ‘OneVsRestClassifier’ for classifying
- Parameter: Specified the parameters to be tuned using grid search
- Parameters for TF-IDF vectorization (max_df, ngram_range, min_df) and same for CountVectorizer (Specific Parameters with different classifiers)
- Logisctic regression classifier (c, which is the inverse regularization strength) , Naive Bayes classifier (alpha), Linear Support Vector Classifier, parameters include 'C' and 'class_weight'.
- GridSearchCV: Object with the pipeline, parameters, 2-fold cross-validation, 3 parallel jobs for processing, and verbose output with level 10
- Retrieves the best classifier
- Makes predictions on the test data using the best classifier

Evaluation Metric – F1 Score

Why f1 ?

- Accuracy can be assessed in two ways: 1.) by considering correct predictions only if all genres match, or 2.) by calculating accuracy based on the number of correct genre predictions out of all predictions made for each movie.
- First method penalizes heavily for any single incorrect genre prediction out of 27. Second method may not be optimal due to dataset imbalance.
- Precision and recall are better metrics for assessing performance, particularly in the case of imbalanced data.
- Precision focuses on the accuracy of positive class predictions, indicating the proportion of correctly identified positive cases among all cases identified as positive.

For each Genre

$$\text{Precision(Genre = Crime)} = \frac{\text{No. of movies 'correctly' identified as Crime Genre}}{\text{Total no. of movies that have been identified as Crime Genre)}}$$

$$\text{Recall (Genre = Crime)} = \frac{\text{No. of movies 'correctly' identified as Crime Genre}}{\text{Total no. of Crime Genre movies in the data set}}$$

$$\text{F1 score (Crime)} = \frac{2 * \text{Precision(crime)} * \text{Recall(crime)}}{\text{Precision(crime)} + \text{Recall(crime)}}$$

- **Overall F1 Score = Weighted Average of individual Genre F1 Score**

Output: BR+ Tf-Idf + LR

Total we had 18 combinations from which the best parameter set is given below:

```
Best parameters set:
[('tfidf', TfidfVectorizer(max_df=0.5, ngram_range=(1, 2))), ('clf', OneVsRestClassifier(estimator=LogisticRegression(C=1), n_j
obs=1))]
```

	Precision	Recall	F1-Score	Support
Action	0.95	0.28	0.44	4321.0
Adult	0.00	0.00	0.00	11.0
Adventure	0.89	0.21	0.34	3496.0
Animation	0.92	0.31	0.47	3333.0
Biography	0.00	0.00	0.00	354.0
Comedy	0.85	0.55	0.67	7320.0
Crime	0.88	0.54	0.67	4453.0
Documentary	0.80	0.35	0.49	1863.0
Drama	0.86	0.78	0.82	11067.0
Family	0.93	0.29	0.44	4173.0
Fantasy	0.93	0.15	0.26	2643.0
Game-Show	0.89	0.38	0.54	450.0
History	0.88	0.07	0.14	623.0
Horror	1.00	0.02	0.04	854.0
Music	0.88	0.18	0.29	654.0
Musical	0.00	0.00	0.00	182.0
Mystery	0.84	0.37	0.51	4114.0
News	0.93	0.40	0.56	681.0
Reality-TV	0.90	0.43	0.58	1748.0
Romance	0.93	0.51	0.66	4581.0
Sci-Fi	0.92	0.29	0.45	3055.0
Short	0.00	0.00	0.00	142.0
Sport	0.86	0.08	0.14	426.0
Talk-Show	0.89	0.48	0.63	809.0
Thriller	0.82	0.19	0.31	3254.0
War	0.96	0.06	0.11	388.0
Western	0.93	0.25	0.40	445.0
Avg/Total	0.88	0.42	0.54	65440.0

Inference:

- We have calculated precision, recall and f-1 score for all the genres individually to understand how accurately each genre is being predicted
- Overall our precision value is 0.88, Recall - 0.42 and F-1 score is 0.54
- For this algorithm we obtained least F-1 score
- Worst performing genres are Adult, Biography, Short and Musical

Output: BR+ Tf-Idf + MNB

Total we had 18 combinations from which the best parameter set is gicen below:

```
Best parameters set:
[('tfidf', TfidfVectorizer(max_df=0.5, min_df=5, ngram_range=(1, 2))), ('clf', OneVsRestClassifier(estimator=MultinomialNB(alpha=0.01)))]
```

Inference:

- Overall our precision value is 0.90, Recall - 0.53 and weighted F-1 score is 0.65
- Best performing genres are Drama, Crime, Game-Show and Romance

Applying best classifier on test data:

	Precision	Recall	F1-Score	Support
Action	0.95	0.42	0.59	4321.0
Adult	0.00	0.00	0.00	11.0
Adventure	0.91	0.43	0.59	3496.0
Animation	0.94	0.52	0.67	3333.0
Biography	0.80	0.08	0.14	354.0
Comedy	0.89	0.60	0.71	7320.0
Crime	0.91	0.59	0.72	4453.0
Documentary	0.75	0.50	0.60	1863.0
Drama	0.87	0.80	0.84	11067.0
Family	0.95	0.42	0.58	4173.0
Fantasy	0.92	0.36	0.52	2643.0
Game-Show	0.90	0.60	0.72	450.0
History	0.74	0.30	0.42	623.0
Horror	0.97	0.14	0.25	854.0
Music	0.90	0.33	0.48	654.0
Musical	0.95	0.12	0.21	182.0
Mystery	0.87	0.43	0.57	4114.0
News	0.85	0.58	0.69	681.0
Reality-TV	0.87	0.51	0.65	1748.0
Romance	0.94	0.58	0.71	4581.0
Sci-Fi	0.94	0.44	0.60	3055.0
Short	0.83	0.04	0.07	142.0
Sport	0.89	0.28	0.42	426.0
Talk-Show	0.82	0.64	0.72	809.0
Thriller	0.88	0.32	0.47	3254.0
War	0.93	0.30	0.46	388.0
Western	0.95	0.44	0.61	445.0
Avg/Total	0.90	0.53	0.65	65440.0

Output: BR+ Tf-Idf + SVC

Total we had 18 combinations from which the best parameter set is gicen below:

```
Best parameters set:
{'clf__estimator__C': 10, 'clf__estimator__class_weight': 'balanced', 'tfidf__max_df': 0.25, 'tfidf__min_df': 1, 'tfidf__ngr
am_range': (1, 2)}
Best model performance:
0.2129021319883016
```

Inference:

- Overall our precision value is 0.80 and Recall - 0.68 and
- Weighted F-1 score is 0.76
- This is the best model
- Genres with highest precision value are Musical , Action, Animation with value 0.98 and 0.92 respectively.

Applying best classifier on test data:

	Precision	Recall	F1-Score	Support
Action	0.92	0.64	0.75	4321.0
Adult	0.00	0.00	0.00	11.0
Adventure	0.91	0.58	0.70	3496.0
Animation	0.92	0.70	0.80	3333.0
Biography	0.88	0.13	0.23	354.0
Comedy	0.85	0.73	0.79	7320.0
Crime	0.88	0.76	0.81	4453.0
Documentary	0.77	0.64	0.70	1863.0
Drama	0.88	0.86	0.87	11067.0
Family	0.90	0.65	0.76	4173.0
Fantasy	0.92	0.54	0.68	2643.0
Game-Show	0.92	0.70	0.79	450.0
History	0.82	0.37	0.51	623.0
Horror	0.92	0.31	0.46	854.0
Music	0.91	0.54	0.68	654.0
Musical	0.98	0.25	0.40	182.0
Mystery	0.86	0.63	0.73	4114.0
News	0.92	0.66	0.77	681.0
Reality-TV	0.85	0.71	0.78	1748.0
Romance	0.92	0.72	0.81	4581.0
Sci-Fi	0.91	0.63	0.74	3055.0
Short	0.94	0.11	0.20	142.0
Sport	0.90	0.46	0.61	426.0
Talk-Show	0.89	0.76	0.82	809.0
Thriller	0.89	0.51	0.64	3254.0
War	0.94	0.47	0.63	388.0
Western	0.91	0.67	0.77	445.0
Avg/Total	0.89	0.68	0.76	65440.0

Output: BR+ CV + SVC

Total we had 18 combinations from which the best parameter set is gicen below:

```
Best parameters set:
[('cvec', CountVectorizer(max_df=0.5, min_df=2, ngram_range=(1, 2))), ('clf', OneVsRestClassifier(estimator=LinearSVC(C=1, clas
s_weight='balanced')))]
```

Inference:

- Overall our precision value is 0.80 and Recall - 0.68
- Weighted F-1 score is 0.76
- Best performing genre is Drama with F-1 score 0.87

Applying best classifier on test data:

	Precision	Recall	F1-Score	Support
Action	0.88	0.59	0.71	4321.0
Adult	0.00	0.00	0.00	11.0
Adventure	0.87	0.53	0.66	3496.0
Animation	0.89	0.67	0.76	3333.0
Biography	0.75	0.16	0.27	354.0
Comedy	0.81	0.71	0.76	7320.0
Crime	0.88	0.68	0.76	4453.0
Documentary	0.75	0.60	0.67	1863.0
Drama	0.89	0.78	0.83	11067.0
Family	0.87	0.62	0.72	4173.0
Fantasy	0.87	0.50	0.64	2643.0
Game-Show	0.93	0.66	0.77	450.0
History	0.78	0.36	0.50	623.0
Horror	0.80	0.30	0.44	854.0
Music	0.88	0.51	0.65	654.0
Musical	0.88	0.24	0.38	182.0
Mystery	0.86	0.55	0.67	4114.0
News	0.90	0.65	0.75	681.0
Reality-TV	0.80	0.68	0.73	1748.0
Romance	0.90	0.67	0.77	4581.0
Sci-Fi	0.88	0.57	0.69	3055.0
Short	0.81	0.12	0.21	142.0
Sport	0.83	0.43	0.57	426.0
Talk-Show	0.88	0.71	0.79	809.0
Thriller	0.84	0.46	0.59	3254.0
War	0.90	0.42	0.57	388.0
Western	0.86	0.60	0.71	445.0
Avg/Total	0.86	0.63	0.72	65440.0

Performance metrics

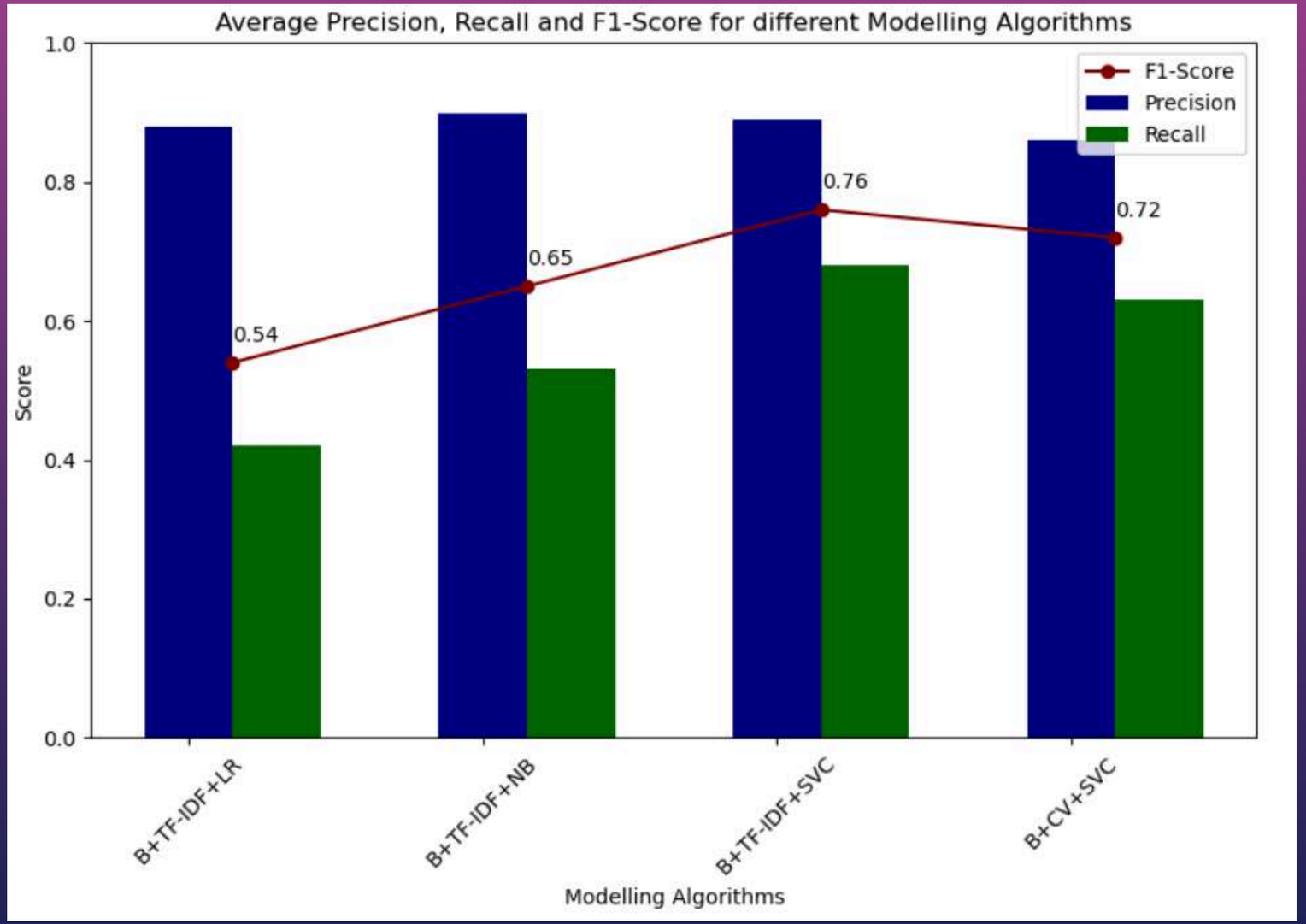
Model	Precision	Recall	F-1 score
BR+ Tf-Idf + LR	0.88	0.42	0.54
BR+ Tf-Idf + MNB	0.90	0.53	0.65
BR+ Tf-Idf + SVC	0.89	0.68	0.76
BR+ CV + SVC	0.86	0.63	0.72

Observations:

- BR + Tf-Idf + svc is the best combination with the f-1 score of 0.76
- the least performer is br+ tf-idf+lr with f-1 score of only 0.54

Comparison of Model Performance: Precision, Recall, and F1-Score

The bar chart summarizes the average Precision, Recall, and F1-score across all models used.



Challenges

1. The dataset was **imbalanced** as all genres were not uniformorly distributed which bias the model towards the majority classes and hence, reducing its ability to accurately predict minority classes.

   How we tackled ??

- Assign **higher weights to minority classes** during model training to penalize misclassifications of minority classes more heavily

2. The predicted genres by the model do not exactly match the true genres of a movie

   how we tackled:
   Threshold Adjustment:
   - Instead of predicting a binary outcome for each genre (e.g., presence or absence), we predicted the probability of each genre. Adjusting the prediction threshold for each genre based on the confidence level can help balance precision and recall, allowing for more flexible predictions.
   - Used evaluation metrics that account for partial correctness such as F1-score

Different Probability for all genres

Action	0.105594
Adult	0.000521
Adventure	0.087402
Animation	0.097010
Biography	0.011818
Comedy	0.289008
Crime	0.129025
Documentary	0.102565
Drama	0.391581
Family	0.131440
Fantasy	0.060609
Game-Show	0.017450
History	0.022714
Horror	0.021938
Music	0.024208
Musical	0.005077
Mystery	0.102616
News	0.033662
Reality-TV	0.105261
Romance	0.163609
Sci-Fi	0.073878
Short	0.004932
Sport	0.016596
Talk-Show	0.044823
Thriller	0.075533
War	0.012006
Western	0.023559

dtype: float64

- Each genre has a threshold value associated with it. For instance, the threshold for Action is approximately 0.106
- These thresholds likely represent the minimum probability required for a movie to be classified under each respective genre based on its plot description.
- For example, a movie needs to have at least a 10.6% predicted probability for the Action genre for it to be classified as such.
- Genres like Drama have higher thresholds (around 0.392), indicating they are more common or perhaps the model requires higher confidence to classify a movie as a drama.

FUTURE SCOPE:

- **Capture inter label dependencies:** Implement hierarchical classification to capture hierarchical relationships between genres, or use techniques like multi-label feature selection to identify discriminative features for each genre while accounting for overlap.

- Could explore advanced techniques such as deep learning architectures (e.g., neural networks) and ensemble methods to further improve the accuracy and robustness of multi-label movie genre classification models.

CONCLUSION:

Our project addressed the demand for personalized content recommendations in streaming platforms by developing a multi-label genre classification system for movies.
Through thorough preprocessing and model optimization, we achieved a notable improvement, obtaining a weighted F-1 score of 0.76.
This signifies enhanced accuracy in assigning multiple genre labels to movies, ultimately improving the system's ability to offer tailored content suggestions to users.
As we look ahead, continued refinement using advanced techniques will be crucial in capturing the nuanced complexities of movie narratives and user preferences, ensuring an enriched streaming experience for audiences globally.

OURTEAM

Anshika Singh

Harsh Mishra

Rohit Singh

REFERENCES

- A. M. Ertugrul and P. Karagoz, "Movie Genre Classification from Plot Summaries Using Bidirectional LSTM", 12th International Conference on Semantic Computing (ICSC), 2018
- R. B. Mangolin, R. M. Pereira, A. S. Britto, C. N. Silla, V. D. Feltrim, D. Bertolini, et al., "A multimodal approach for multi-label movie genre classification", Multimedia Tools and Applications, 2020
- R. Vidiyala, “How to build a movie recommendation system - towards data science,” Medium, Aug. 09, 2022. [Online]. Available: https://towardsdatascience.com/how-to-build-a-movie-recommendation-system-67e321339109#:~:text=A)%20Content%2DBased%20Movie%20Recommendation%20Systems,-Content%2Dbased%20methods&text=Using%20this%20type%20of%20recommender,the%20same%20actor%2C%20or%20both.
- “What is RNN? - Recurrent Neural Networks Explained - AWS,” Amazon Web Services, Inc. https://aws.amazon.com/what-is/recurrent-neural-network/#:~:text=A%20recurrent%20neural%20network%20(RNN,a%20specific%20sequential%20data%20output.


THANKYOU !